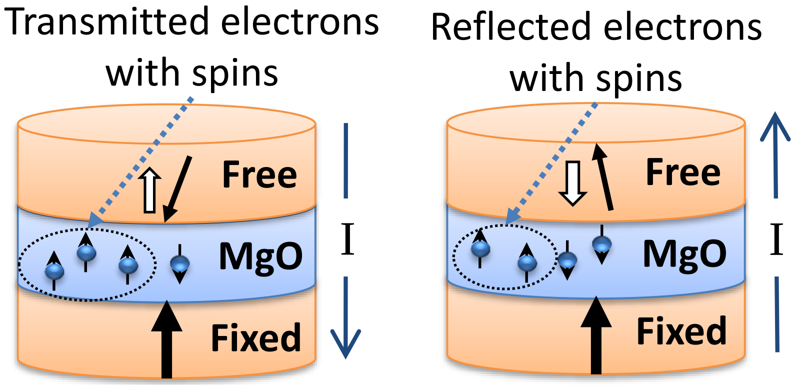
Shaodi Wang, Hochul Lee, Pedram Khalili, Cecile Grezes, Kang L. Wang and Puneet Gupta

University of California, Los Angeles

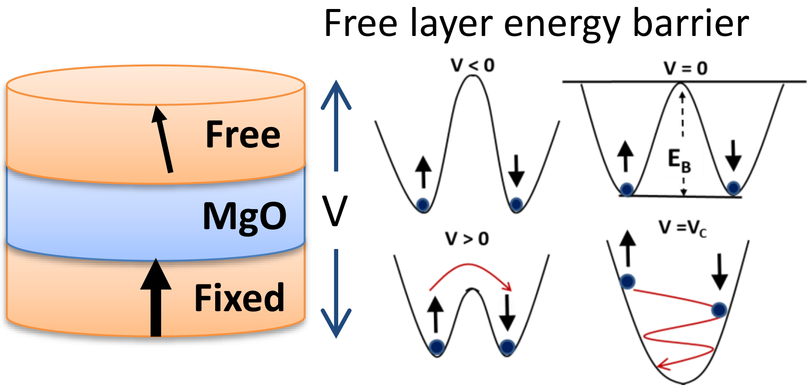
VARIATION MONITOR-ASSISTED ADAPTIVE MRAM WRITE

Write mechanism of STT-RAM and MeRAM

Spin-torque transfer magnetic tunnel junction (STT-MTJ)


Transmitted electrons with spins
Reflected electrons with spins
Free
MgO
Fixed
I
Free
MgO
Fixed
I

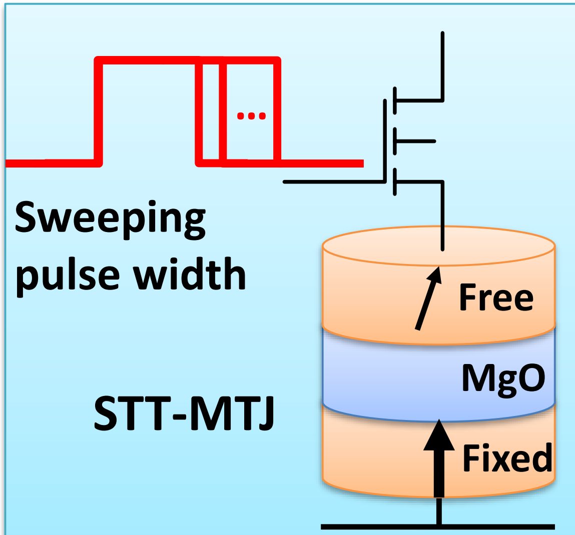
Voltage-control magnetic tunnel junction (VC-MTJ)



- STT-MTJ write
  - **Bi-directional current-driven**
  - Critical current density (J_C)
  - Deterministic write
  - **Slow (5~10ns)**
  - **High power (0.2pJ~1 *pJ/bit*) due to low MTJ resistance (1k-10k Ω)**

- VC-MTJ write
  - **Uni-directional voltage-driven**
  - Critical voltage (V_C)
  - **Non-deterministic write (leads to write errors)**
  - Fast(~1ns)
  - Low power (10~50 *fJ/bit*) due to high MTJ resistance (20k-200k Ω)

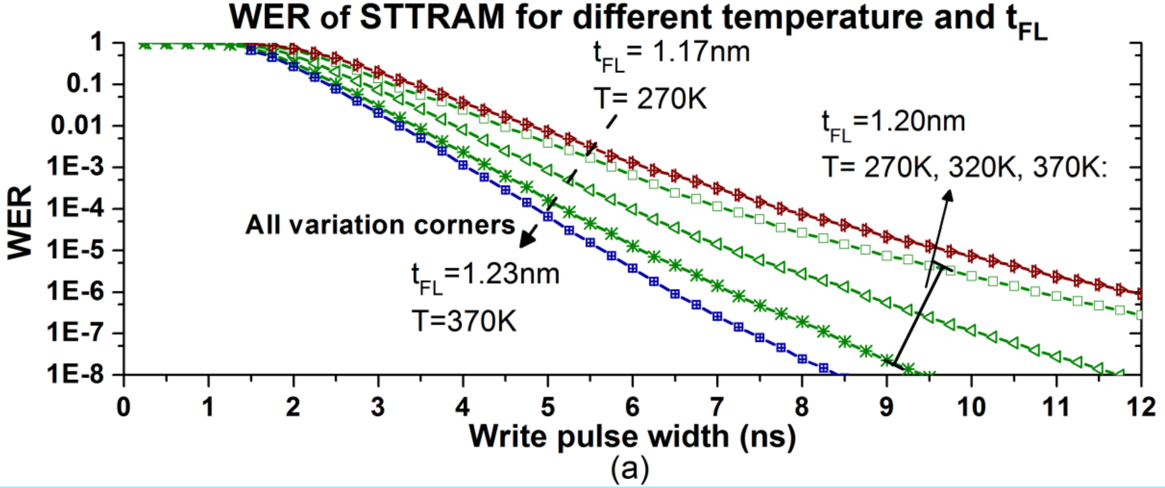
MRAM write error rate (WER) under variation



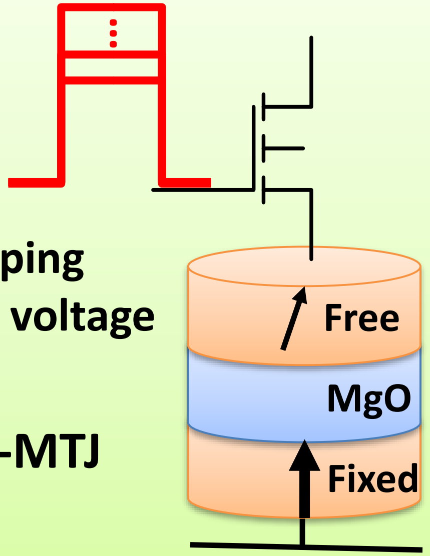
Sweeping pulse width

STT-MTJ

Spin-transfer torque RAM (STT-RAM)



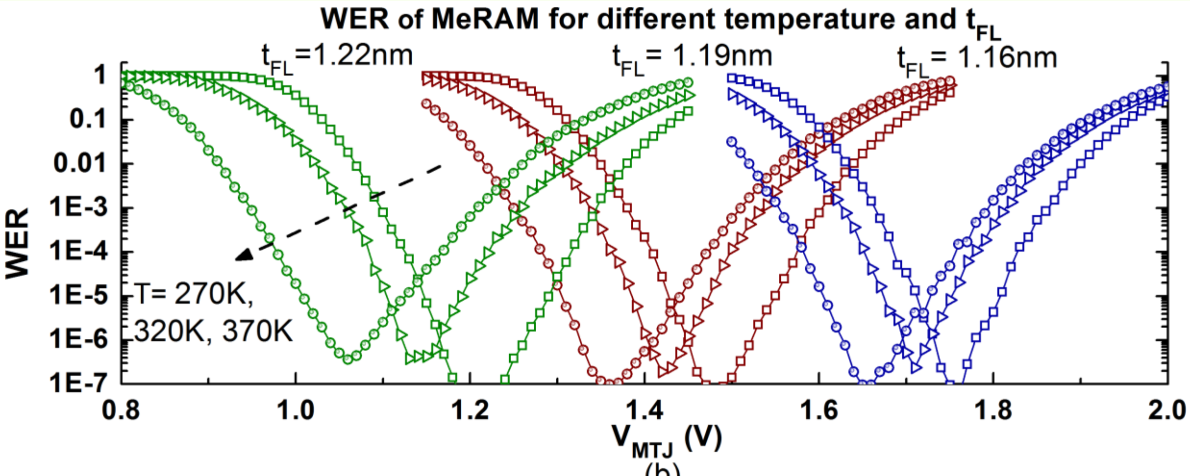
(a)



Sweeping pulse voltage

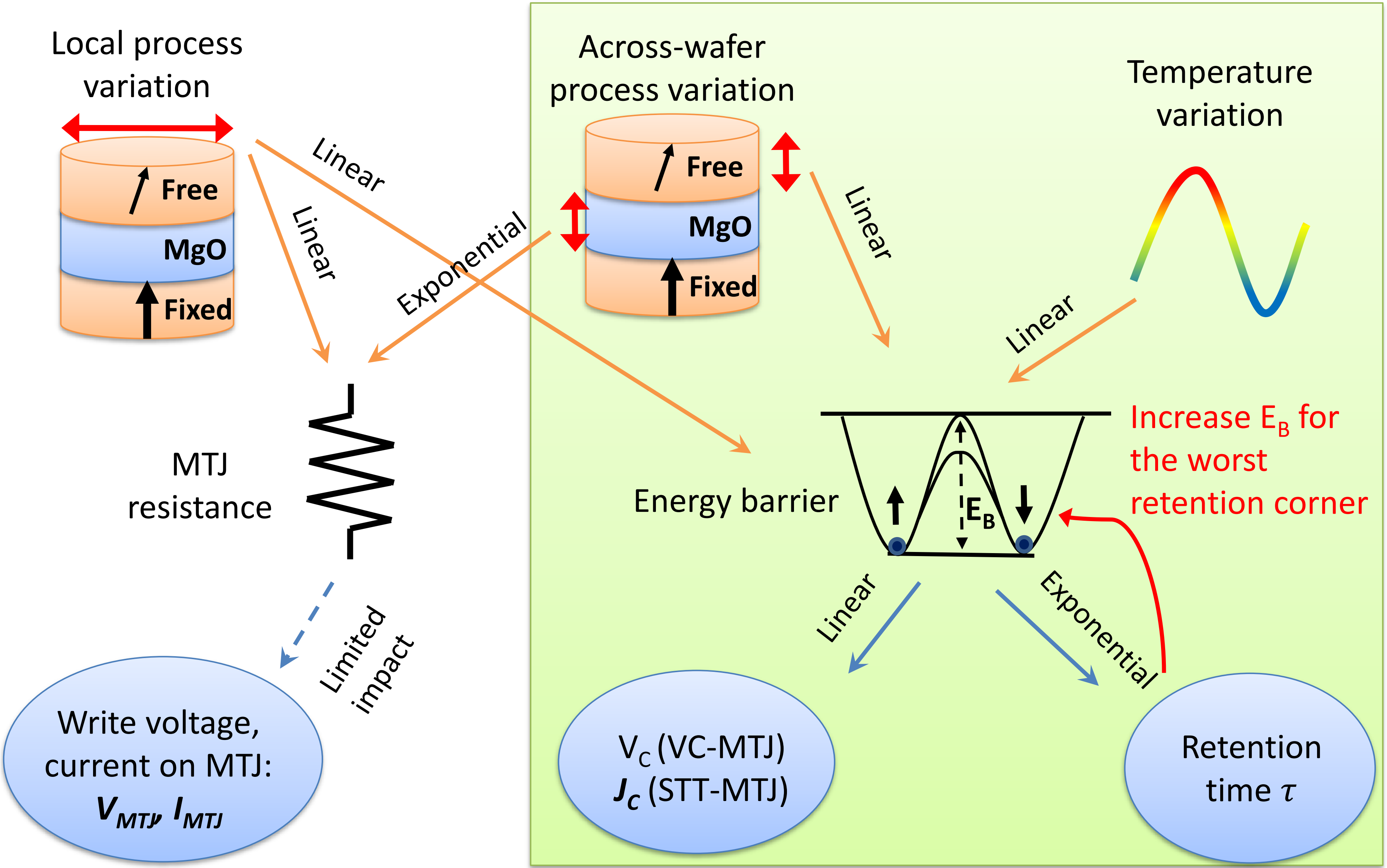
VC-MTJ

Magnetoelectric RAM (MeRAM)



(b)

MRAM write under variation



Local process variation

Across-wafer process variation

Temperature variation

MTJ resistance

Energy barrier

Increase E_B for the worst retention corner

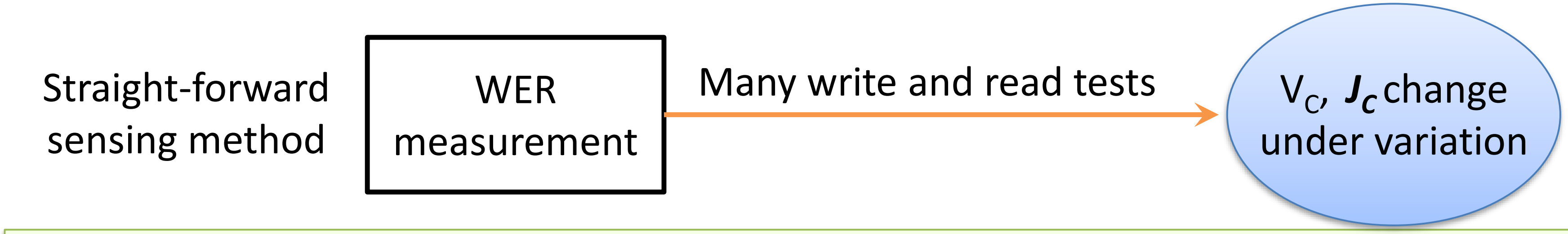
Write voltage, current on MTJ: $\boldsymbol{V_{MTJ}, I_{MTJ}}$

V_C (VC-MTJ)
$\boldsymbol{J_C}$ (STT-MTJ)

Retention time τ

Sensing write behavior change under variation

30°C changes WER from 10^{-6} to 10^{-4} →High energy and long delay

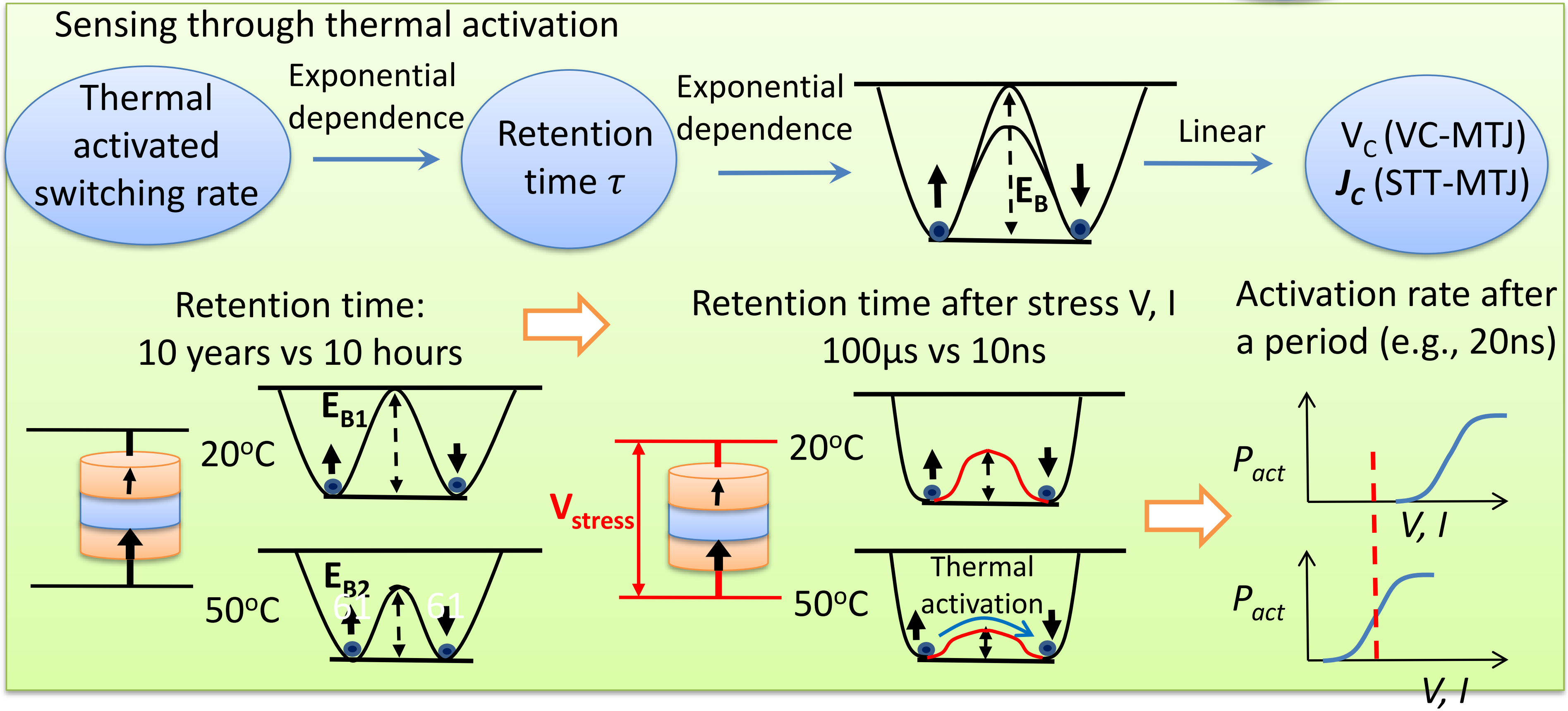


Straight-forward sensing method

WER measurement

Many write and read tests

V_C, $\boldsymbol{J_c}$ change under variation



Sensing through thermal activation

Thermal activated switching rate

Exponential dependence

Retention time τ

Exponential dependence

Linear

V_C (VC-MTJ)

$\boldsymbol{J_c}$ (STT-MTJ)

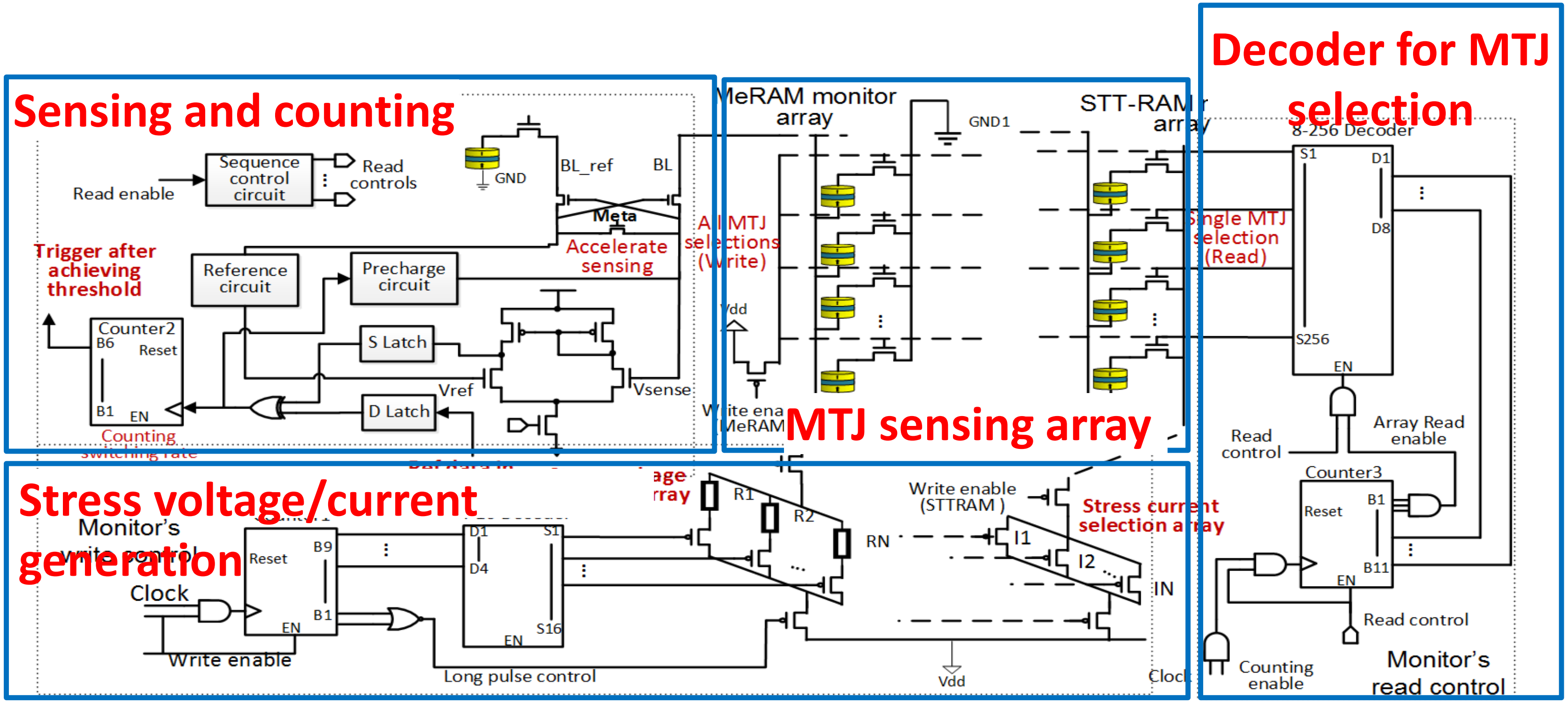
Retention time:

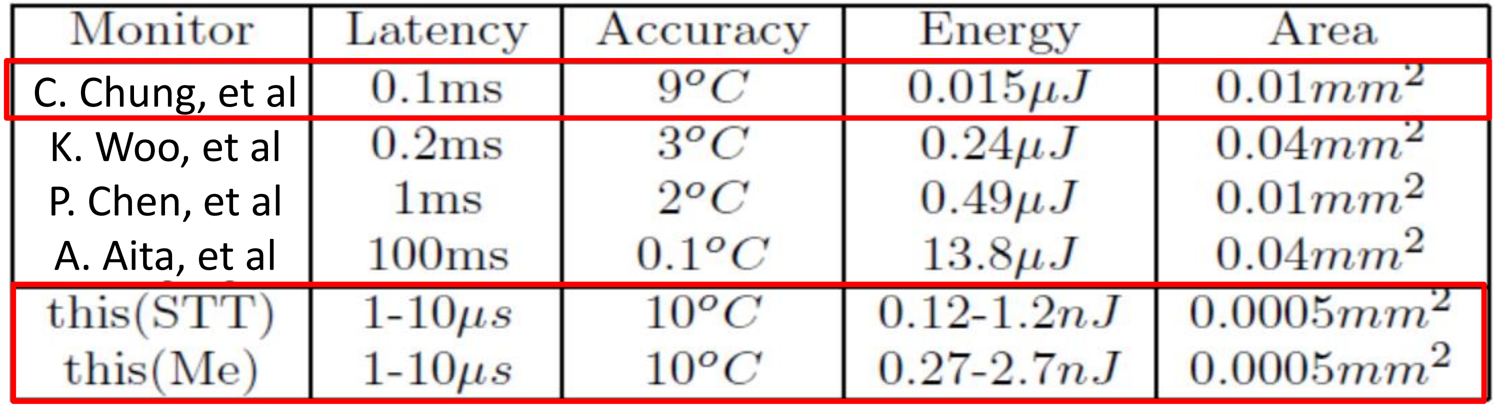
10 years vs 10 hours

Retention time after stress V, I

100µs vs 10ns

Activation rate after a period (e.g., 20ns)

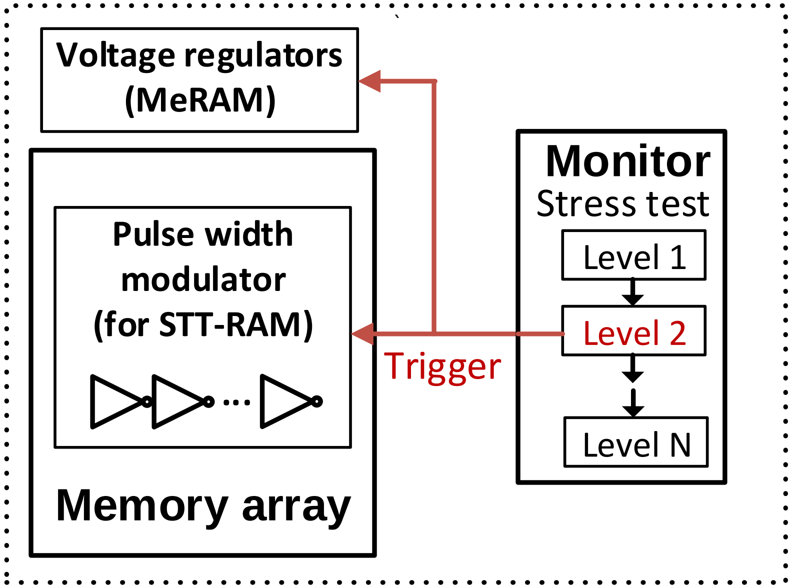




Monitor	Latency	Accuracy	Energy	Area
C. Chung, et al	0.1ms	9^oC	$0.015\mu J$	$0.01mm^2$
K. Woo, et al	0.2ms	3^oC	$0.24\mu J$	$0.04mm^2$
P. Chen, et al	1ms	2^oC	$0.49\mu J$	$0.01mm^2$
A. Aita, et al	100ms	0.1^oC	$13.8\mu J$	$0.04mm^2$
this(STT)	$1\text{-}10\mu s$	10^oC	$0.12\text{-}1.2nJ$	$0.0005mm^2$
this(Me)	$1\text{-}10\mu s$	10^oC	$0.27\text{-}2.7nJ$	$0.0005mm^2$

Application of the variation monitor - adaptive write

- Dynamically select optimal pulses for multiple-write[1]
  - Write latency variation minimization
    - Three write pulse choices are enough
    - 1.2X for 1-MB STT-RAM write latency improvement
    - 2.4X for 1-MB MeRAM write latency improvement



Voltage regulators (MeRAM)

Pulse width modulator (for STT-RAM)

Memory array

Trigger

Monitor

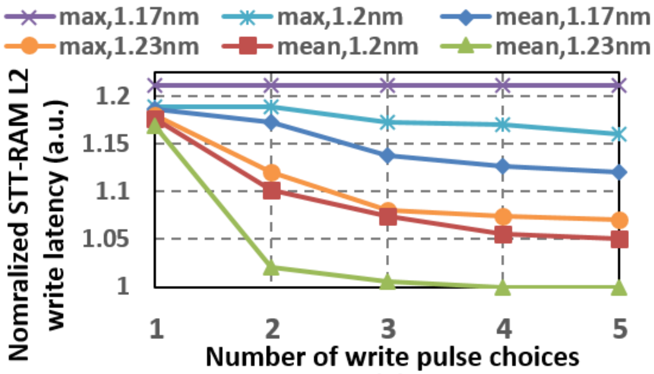
Stress test

Level 1

Level 2

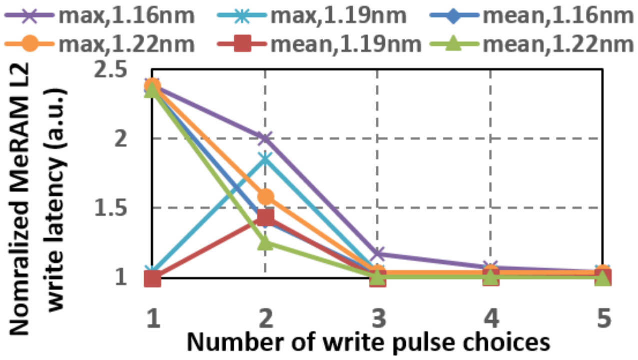
Level N

STT-RAM



(a)

MeRAM

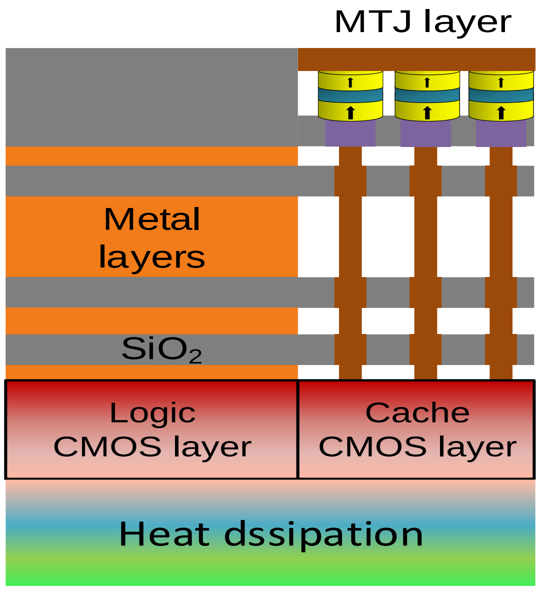


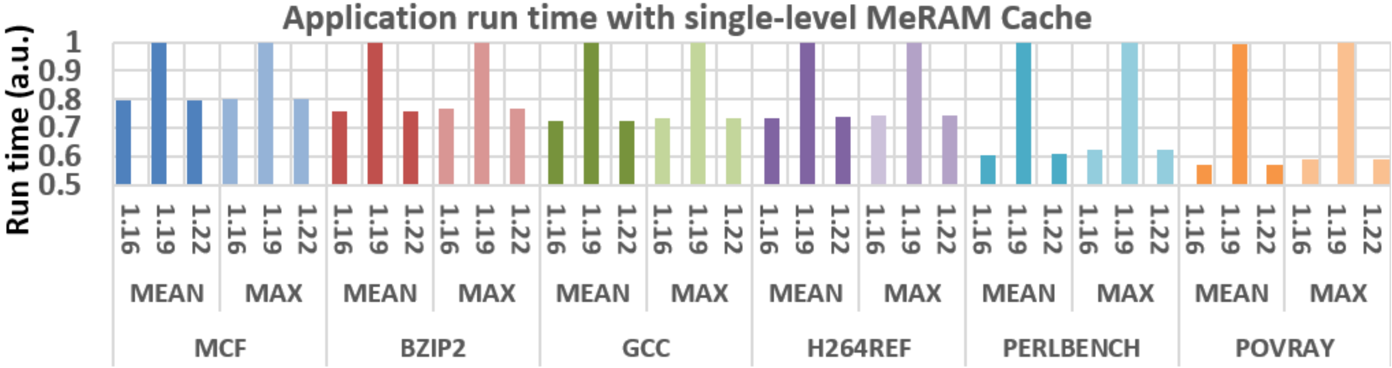
(b)

[1] H. Lee et al. TMAG (2015).

Evaluation of adaptive write

- Experimental setup:
  - 32nm Single-core X86, 8-MB universal MRAM cache
- Simulations
  - MTJ switching simulation (experimentally verified physical models)
  - Circuit simulation (SPICE and NVSIM)
  - Architecture simulation (gem5)
  - Thermal simulation (Hotspot)
  - Power simulation (CACTI)
- 1.7X and 1.1X application run time improvement for processor with MeRAM and STT-RAM





(b)

Conclusion

- The proposed variation monitor can sense combined wafer-level process and temperature variation
  - ***10X faster, 5X energy-efficient, and 20X smaller than conventional 65nm temperature monitor with same accuracy***
- Adaptive write scheme dynamically selects optimized write pulse through variation monitoring
  - MeRAM receives more benefit than STT-RAM
    - ***2.4X and 1.2X cache speed improvement for MeRAM and STT-RAM***
    - ***MeRAM suffers from more variation impact***
    - STT-RAM without multiple-write is expected to see much more improvement in both power and latency (future work)
    - 1.7X application run time reduction for processor with MeRAM cache
    - 1.1X application run time reduction for processor with STT-RAM cache
- **Thank you for your attention**